2006

Virtual evidence: analyze the footsteps of your users

Win Shih

Virtual Evidence: Analyze the Footsteps of Your Users

Win Shih

ABSTRACT. This paper presents a study of Web Crawler activities based upon Web access logs from the Web site of an academic library. It further compares crawler behavior with that of regular human visitors. The results provide practical insights and foster a culture of evidence-based practice for better managing network-based resources and maintaining a reliable IT infrastructure. doi:10.1300/J186v06n04_04 *[Article copies available for a fee from The Haworth Document Delivery Service: 1-800-HAWORTH. E-mail address: <docdelivery@haworthpress.com> Website: <http://www.HaworthPress.com>*

KEYWORDS. Web crawler, Internet robot, Web server management, network management, Web access log analysis

INTRODUCTION

Inside the networked academy of today, the library's Web site is the portal to ever-burgeoning electronic resources, support and services. Understanding the nature and characteristics of how your site is utilized becomes crucial for ensuring and improving the quality of services your

Win Shih is Assistant Professor and Head of Systems and Databases, Denison Memorial Library, University of Colorado and the Health Sciences Center, Denver, CO (E-mail: win.shih@uchas.edu).

Journal of Hospital Librarianship, Vol. 6(4) 2006
Available online at http://jhspl.haworthpress.com

doi:10.1300/J186v06n04_04

prized patrons receive upon virtual demand. Monitoring Web site usage and server workload not only assures 100% uptime performance, but concomitantly assists in improving Web site design, systems performance, and resource utilization, in conjunction with future hardware and infrastructure capacity planning.

The intensifying imperative of mega-powerful, Web-driven search engines such as Google and Yahoo, combined with specialized "Internet Search Assistants," have contributed to an exponential increase in the population, as well as variety, of software agents, commonly known as Internet robots or simply "bots." These agents are software programs that traverse the Internet in an automated, methodical manner, gathering or "harvesting" the content of each Web site they visit, including those of academic libraries and hospitals. Bots tend to consume a considerable amount of system resources and network bandwidth from the sites they frequent, at the expense of regular site users and usage. Additionally, they generate concerns regarding unauthorized content collection, privacy, not to mention a host of other security-centric matters.

This paper reports a study identifying autonomous software agents and their impact on a library's Web site based on Web access logs. Furthermore, it characterizes behavior of the major software agents.

COLLECTION OF DATA

This study employs the Web access logs recorded on the Web server at University of Colorado Health Sciences Center Library (http://denison.uchsc.edu/). The data covers a four-month period, from March 27 through July 26, 2005. Table 1 provides a summary of the characteristics of the raw data. "Analog" (http://www.analog.cx/), a popular and free software program which analyzes and summarizes Web access log files, was used to process 122 daily log files.

CRAWLER ACTIVITIES

On average, Web crawlers daily paid 1,144 *Page* visits and 1,346 Requests during the period of investigation. Figure 1 shows the distribution of *Page* requests during different hours of the day. To our surprise, the line chart shows a spike of heavy usage at midnight. After further examining the log entries, we learned that the University's Information Technology Services ran an enterprise search engine soft-

TABLE 1. A Summary of Access Log Characteristics (the Raw Data)

Log Period	**3/27/05-7/26/05**
Log duration (days)	122
Log size (MB)	762
Total *Requests* (total number of files downloaded,including graphics)	4,820,709
Total requests for *Pages* (total Web pages requested)	701,261
Average daily *Requests*	39,514
Average Web crawler daily *Requests*	1,346
Average daily requests for *Pages*	5,748
Average Web crawler daily requests for *Pages*	1,144

FIGURE 1. *Page* Requests During Different Hours

ware package, called *Ultraseek* early every morning to index all Web servers on campus. The information collected by Ultraseek crawlers is used to construct the University Web site's search engine.

To better understand the actual Web crawler impact outside the context of our unique, localized situation, the entries from the University's Ultraseek crawlers were removed from our log analysis. Figure 2 shows the revised distribution of *Page* requests by "hours of the day."

As one can see from the line chart, our Web site traffic starts to increase after 7 a.m. and reaches its peak around noon time. The usage sustains for another three hours before gradually dropping off at 4 p.m. On the other hand, the activities attributed to Web crawlers display a

steady and stable flow regardless of the time of day. Our findings indicate that Web crawlers pay little attention to the activity status of our Web site, nor do they pay heed to the workload of our Web server or network traffic. As a result, Web crawler activities during busiest periods for our Web site put an additional burden upon an already overloaded system and network. Furthermore, they compete with our regular users for system resources and are likely to impair services overall. Not only can our users not fully access our resources, Web crawlers likewise are unable to harvest their content in the most efficient manner. A similar result was reported by Ye et al. (1).

CRAWLER IMPACT

As Table 2 shows, Web crawlers account for less than 2.41% of all *Requests* received by the Web server during the 122-day period. However, when we focus on the number of Web pages viewed by Web

FIGURE 2. *Page* Requests During Different Hours Excluding Ultraseek Crawler Activities

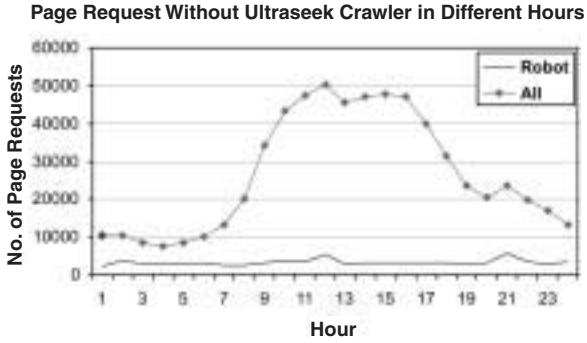

TABLE 2. Contribution of Web Crawlers to Web Server Activity

	All	Web Crawlers	Percentage
Total *Requests* (total number of files downloaded, including graphics), excluding Ultraseek crawler activities	4,758355	101,821	2.41
Total requests for *Pages* (HTML pages), excluding Ultraseek crawler activities	640,236	101,821	12.26

crawlers, the percentage jumps to almost 12.26% of total Web page requests. This prodigious percentage difference between *Requests* and *Page* requests is due to the very nature of browsing behavior by Web crawlers. Search engines or Web site harvesters do not usually index image, PDF, or other non-HTML files, and thus their "crawlers" only target HTML files most of the time.

RESOURCE TYPE CRAWLED

The difference in browsing behavior between Web crawlers and regular Web site visitors can be further observed by comparing the type of files requested by each group, as shown in Table 3. More than 90% of the requests by Web crawlers are for text files, which consist of .html, .htm, and directories. These types of files match the definition of a "*Page* visit." In contrast, only 16% of requests from human visitors are for text-type files. On the other hand, image files (.gif, .jpg, and .pdf files), which are practicably negligible to a Web crawler, account for more than 65% of regular user requests. Our findings match those reported by Dikaiakos et al. (2,3) and Tan and Kumar (4).

TOP CRAWLERS

Web crawlers can be identified by reviewing the user agent field of each log entry. Table 4 lists the top ten crawlers and their frequency of

TABLE 3. Comparison of File Type Requests Between Web Crawlers and Regular Users

File Type	% of Web Crawler Requests	% of Non-Crawler Requests
.html (Hypertext Markup Language)	73.89	8.12
.htm (Hypertext Markup Language)	9.07	0.96
(directories)	7.53	6.85
.txt (Plain text)	5.07	0.21
.gif (GIF graphics)	1.64	42.24
.jpg (JPEG graphics)	1.07	23.23
.pdf (Adobe Portable Document Format)	0.53	0.24
.css (Cascading Style Sheets)	0.14	11.45
.js (JavaScript code)	0.07	5.50

TABLE 4. Top Ten Web Crawlers by User Agent

Rank	Crawler	Requests	Pages	Web Crawler Site	Crawler Type
1	UCHSC Ultraseek	62,354	61,025	http://www.verity.com/products/ultraseek/	Enterprise Search Engine
2	Googlebot	13,997	13,503	http://www.google.com/bot.html	Search Engine
3	Yahoo Slurp	13,445	10,002	http://help.yahoo.com/help/us/ysearch/slurp	Search Engine
4	MSNBot	11,923	10,639	http://search.msn.com/msnbot.htm	Search Engine
5	LookSmart	6,845	6,788	http://www.WISEnutbot.com	Search Engine
6	Nutch (Linux-based Open Source free search engine)	5,765	5,454	http://lucene.apache.org/nutch/bot.html	Search Engine
7	FAST Enterprise Crawler	5,512	5,434	http://www.fastsearch.com	Unknown
8	YahooSeeker	3,898	3,798	http://help.yahoo.com/help/us/shop/merchant/	Shopbot
9	Aipbot	2,737	2,606	http://www.aipbot.com	Unknown
10	FAST MetaWeb Crawler	2,243	2,220	http://www.fastsearch.com	Search Engine

visits. After excluding Ultraseek (our campus search engine crawler), we were not surprised to see that the top three crawlers are from the three major commercial Web search engine vendors: Google, Yahoo, and MSN Search. The combined crawling activities from these three commercial search engines account for 43% of all crawler activity (after excluding crawler activity from the University's Ultraseek bot). We were quite amazed to see "YahooSeeker," a "shopbot," listed among the top crawlers, because our site is not an E-commerce site.

CONCLUSIONS AND IMPLICATIONS FOR LIBRARIES

This paper has explored in-depth the extent of and impact by Web crawlers on the Web site of the University of Colorado Health Sciences Center Library. Analyzing our Web access logs, we have concomitantly

been able to learn more about their behavior. Our results strongly suggest that Web crawlers pay frequent visits, yet give little attention to the activity levels of a Web site. Crawlers possess distinctive behavior as compared with that of regular users. They focus on harvesting text-based information, ignoring image files. Our results also identify the most active top 10 crawlers. The top three crawlers are from the major commercial search engines, Google, Yahoo, and MSN Search.

There are several limitations of our study. First, in data collection the task of isolating Web crawlers from regular and legitimate users is a difficult one. The proliferation of Web crawlers is complicated by some robots, such as spambots, disguised as regular browsers and intentionally not providing identification. In this study, we were not able to identify robots camouflaged as regular users and thus in all probability, underreported actual crawler behavior. Moreover, our primary dataset is derived from only one Web server and, therefore, our results are preliminary, limited by sample size, and cannot be generalized to all library Web sites or Web servers. Our quarterly dataset likewise poses a potential shortfall due to lack of comprehensiveness.

Future studies should further identify the intensity and duration of Web crawler visits, the amount of information in bytes requested by crawlers, and finally-more extensively cross-compare crawler behavior with the behavior of regular users. Furthermore, future studies should include Web access logs from multiple Web servers to eliminate uniqueness of a single Web site, as well as longer time coverage (longitudinal studies) to eliminate possible seasonal effects.

"The use of robots comes at a price," stated Koster (5) when he proposed a voluntary robot exclusion standard eventually adopted by robot writers and Web site administrators. There is a delicate tradeoff between complete accessibility versus total exclusion from your Web site. Armed with a sharper, clearer picture of the behavior patterns of Web crawlers attracted to our site at UCDHSC, we have implemented a combination of measures to adjust security of our network and servers. For our Web server, we specify in the "robots.txt" file exactly which portions of our server are off-limits from crawlers. Moving expired Web pages and files to password-protected folders or unloading them from the server altogether effectively prevents our server from becoming clogged with needless or junk files. For our Integrated Library System, we exclude IP addresses from known Web crawlers so we can maximize the usage of limited user licenses. Regularly monitoring access logs and server performance is the best practice for all library Web site system administrators.

Needless to say perhaps, but nonetheless important to reiterate, the challenge of the library Web site administrator is unquestionably a daunting one–balancing the behavior of "benign bots" with that of your regular library user-base, while all the while keeping out the "bad guys"–spambots, viruses, worms, or "poisonous cyber spiders" or any Web crawler whose intent is malign–harvesting your library Web server's invaluable content sans authorization, getting in without a permission or a "clearance." The very best practice we have learned, through trial, error, tons of hard work simply but complicatedly enough: "follow the middle way" to the best of your ability. Not all bots are bad, just as not all human users with a valid password are beyond reproach.

Managing your library's Web server is as much art as it science. Studying those access logs assiduously, learning everything you can about which bots and crawlers are safe or even useful to you and your system, keeping your "robots.txt" file up-to-date, knowing your users–not just the bots but your human user-base as well.

Received: March 16, 2006
Revised: April 4, 2006
Accepted: April 7, 2006

REFERENCES

1. Ye S, Lu G, Li X. (2004). Workload-aware Web crawling and server workload detection. Network Research Workshop, 18th Asian Pacific Advanced Network Meeting (APAN 2004), July 2004. http://www.csif.cs.ucdavis.edu/~yeshao/apan04.pdf (Feb. 18, 2006).

2. Dikaiakos M, Stassopoulous A, Papageorgiou L. An investigation of Web crawler behavior: characterization and metrics. Comput Commun 2005; 28(8): 808-897.

3. Dikaiakos M, Stassopoulous A, Papageorgiou L. Characterizing crawler behavior from Web server access logs. Lect Notes Comput Sci 2003; 2738: 369-378.

4. Tan P, Kumar V. Discovery of Web robots sessions based on their navigational patterns. Data Min Knowl Discov 2002; 6(1): 9-35.

5. Koster, M. (1995). Robots in the web: threat or treat? www.robotstxt.org/wc/threat-or-treat.html (Feb. 18, 2006).